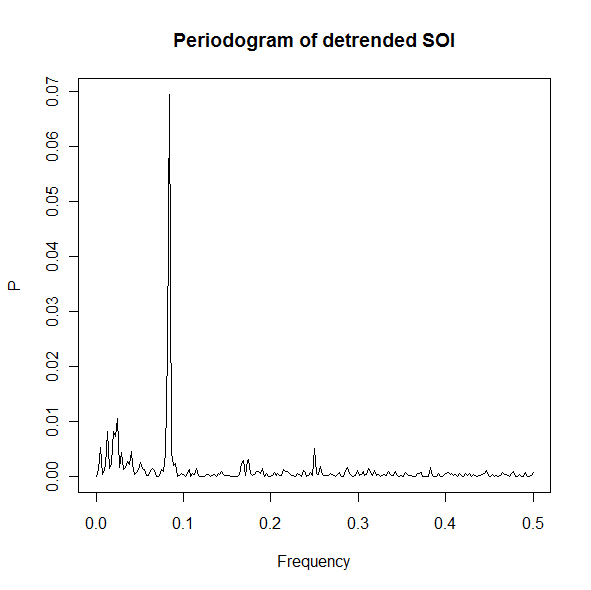
**Stat 510 Week 7 Homework**

**Submit answers in a Word document or a pdf to the designated ANGEL Drop Box in the Drop Boxes folder.**

1. Following is the periodogram of the detrended SOI series that we’ve used a few times in the past. SOI = southern oscillation index, a measure of temperature in the southern hemisphere. The series is n = 452 monthly values. The detrended values are the residuals from a linear regression on time (1 to 452).



**a**. Roughly, at what frequency is the dominant peak? The period of a cycle is the reciprocal of the frequency. About what is the period for the frequency at which the dominant peak occurs? (Keep in mind that this is monthly data and this is a weather measure.)

We see that visually the dominant peak is at 0.08

It makes sense since 1/12 = 0.0833

**b**. Roughly what is the region of frequency range for the second highest peak(s)? Roughly what would be the period(s) associated with these frequencies? (Hint: The El Nino weather phenomenon is believed to affect things about every 3 to 7 years.)

We see that visually the dominant peak is between .02 and .035 (say 0.025)

**2.** Do problem 4.1 on page 255 of the textbook.



For part a, use a modified version of the R code on page 178. To simply increase the sample size from 100 to 128, you need only change 1:100 to 1:128. For part b, use a modification of the code on page 179 or the code below. To print the periodogram between 0 and .5 only, as opposed to between 0 and 1 in Figure 4.2, you may use the following code:

P = abs(2\*fft(x)/128)^2

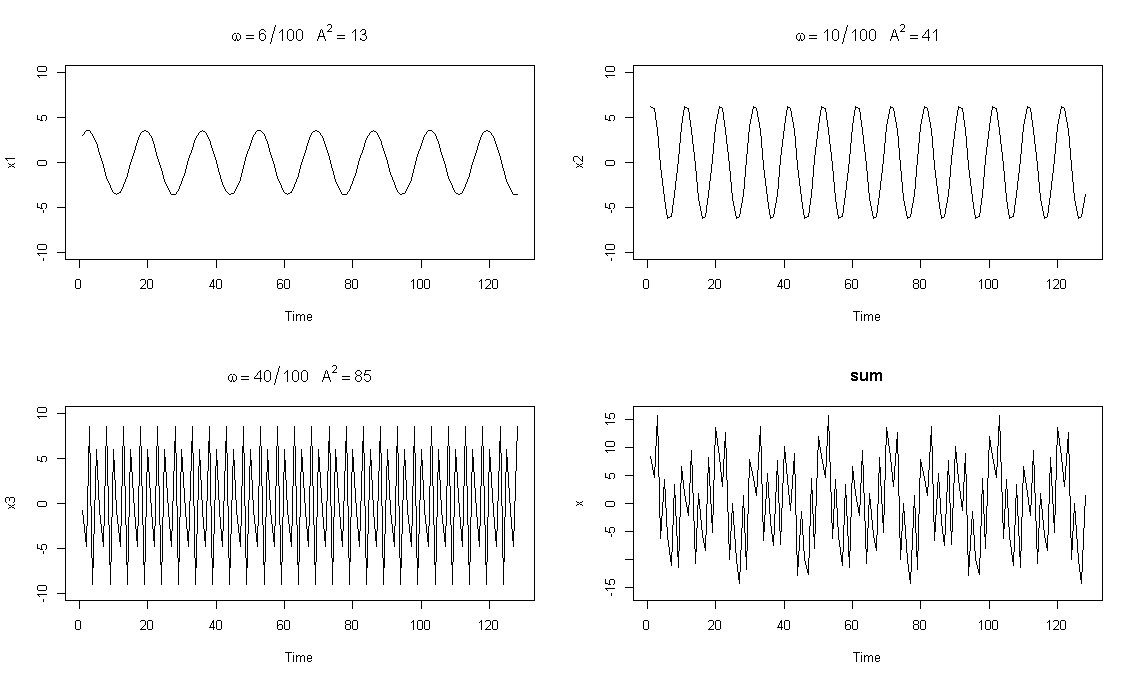
f = 0:64/128

plot(f, P[1:65], type="o", xlab="frequency", ylab="periodogram")

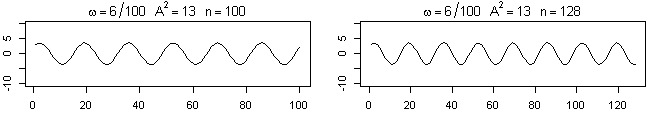
For part c, it’s only necessary to plot the series that is the sum and plot the periodogram of the sum. The command to get the sum will be something like sum = x1+x2+x3+rnorm(100,0,5). The parameters for the rnorm command are sample size, mean, standard deviation. (The notation the authors used in part c gives the mean and the variance.)

For answers to the book parts–

Part a Give plots of the x1, x2, x3 and the sum. To determine how things differ from Example 4.1, focus on the formula for the series rather than the plots.



As far as the difference is concerned, the 2πωjt term has different values. We can say that the frequency will be different. For instance compare:

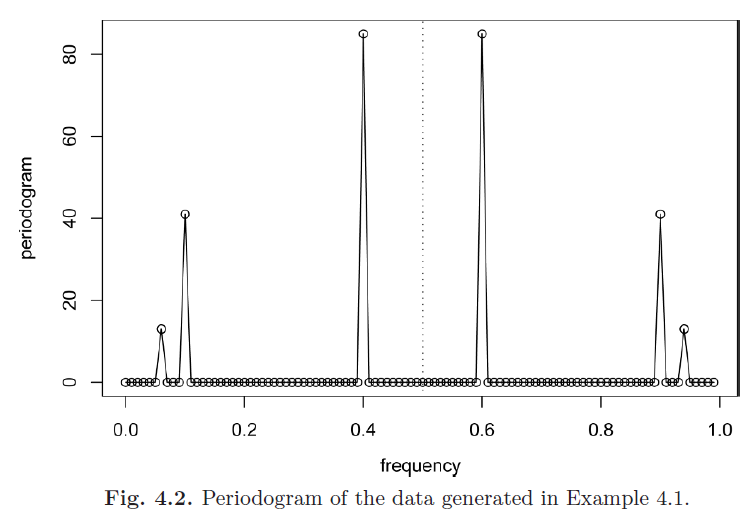


The first image shows almost 6 cycles so freq = 6 / 100

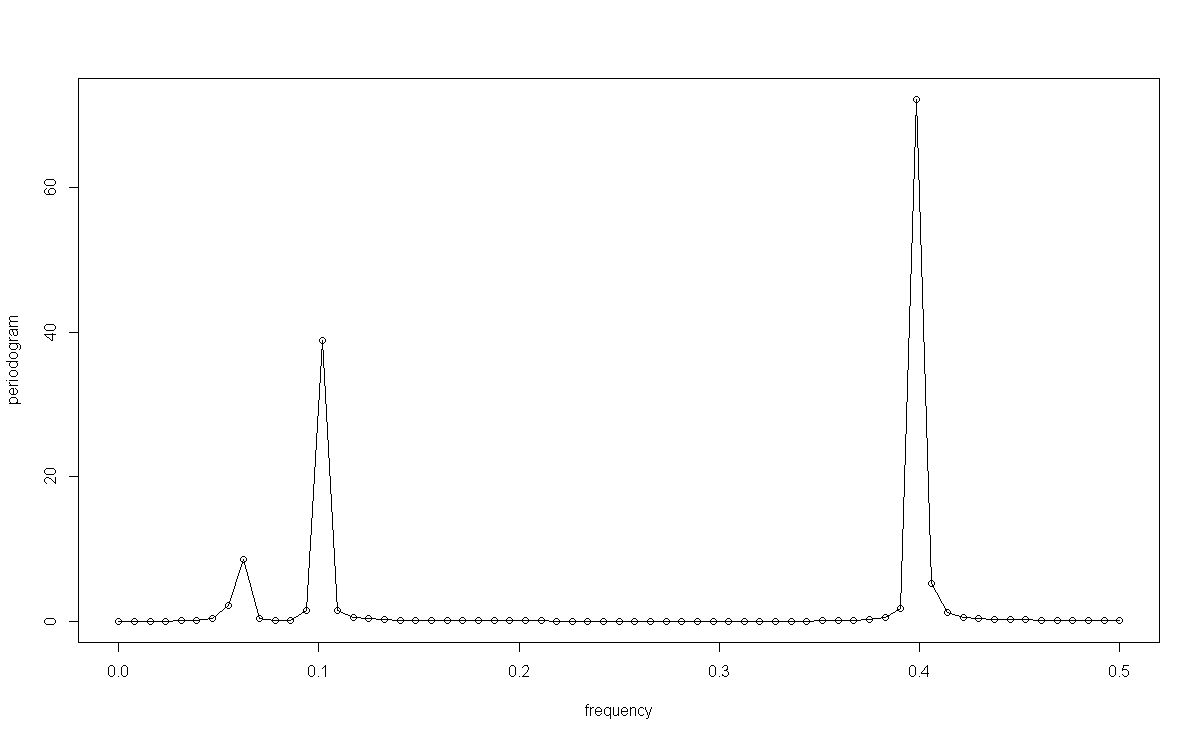
The second image shows 7.5 cycles so freq = 7.5 / 128

Part b. Give the periodogram of the sum and describe the main frequency peaks. What has changed from Figure 4.2?

The original periodogram:



The new periodogram:



Putting them side by side and drawing lines show slight displacement

|  |  |
| --- | --- |
|  |  |
| Old |  |
| New |  |

The main frequency peaks are at: 0.4 (slightly moved away unlike the old), 0.1 (slightly moved away unlike the old), 0.065 (slightly moved away unlike the old)

When we list the values we get exact frequencies: 0.3984375, 0.1015625, 0.0625000

f

[1] 0.0000000 0.0078125 0.0156250 0.0234375 0.0312500 0.0390625 0.0468750

[8] 0.0546875 **0.0625000** 0.0703125 0.0781250 0.0859375 0.0937500 **0.1015625**

[15] 0.1093750 0.1171875 0.1250000 0.1328125 0.1406250 0.1484375 0.1562500

[22] 0.1640625 0.1718750 0.1796875 0.1875000 0.1953125 0.2031250 0.2109375

[29] 0.2187500 0.2265625 0.2343750 0.2421875 0.2500000 0.2578125 0.2656250

[36] 0.2734375 0.2812500 0.2890625 0.2968750 0.3046875 0.3125000 0.3203125

[43] 0.3281250 0.3359375 0.3437500 0.3515625 0.3593750 0.3671875 0.3750000

[50] 0.3828125 0.3906250 **0.3984375** 0.4062500 0.4140625 0.4218750 0.4296875

[57] 0.4375000 0.4453125 0.4531250 0.4609375 0.4687500 0.4765625 0.4843750

[64] 0.4921875 0.5000000

> round(P[1:65],7)

[1] 0.0013557 0.0035626 0.0109761 0.0266241 0.0588234 0.1337226 0.3707554

[8] 2.2207243 **8.5381819** 0.3329350 0.0532147 0.1264216 1.4815151 **38.9092119**

[15] 1.4883467 0.5559387 0.3123064 0.2079457 0.1518081 0.1173670 0.0943501

[22] 0.0780161 0.0658976 0.0565932 0.0492524 0.0433311 0.0384666 0.0344086

[29] 0.0309796 0.0280510 0.0255282 0.0233410 0.0214379 0.0197816 0.0183479

[36] 0.0171238 0.0161093 0.0153191 0.0147881 0.0145805 0.0148062 0.0156500

[43] 0.0174282 0.0206976 0.0264883 0.0368323 0.0561116 0.0950087 0.1847124

[50] 0.4471653 1.7333104 **72.2286650** 5.2307198 1.1965233 0.5718405 0.3584467

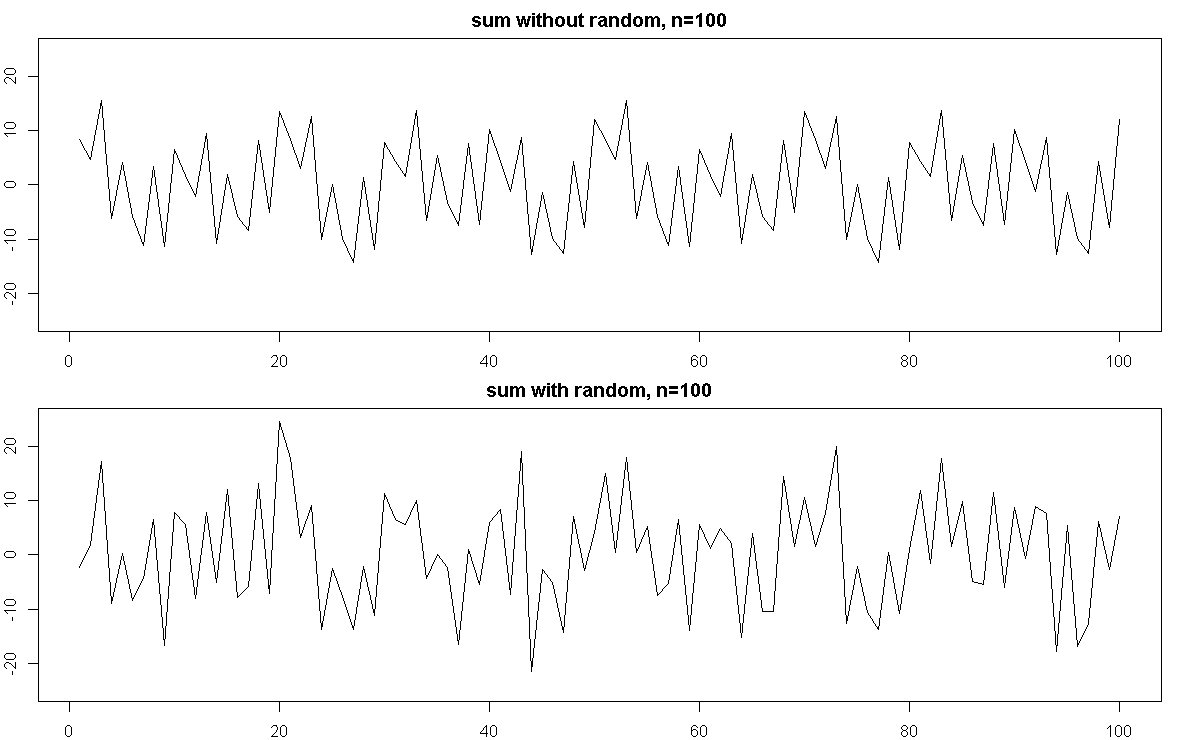
[57] 0.2589509 0.2041721 0.1708619 0.1493536 0.1350325 0.1254834 0.1193721

[64] 0.1159558 0.1148558

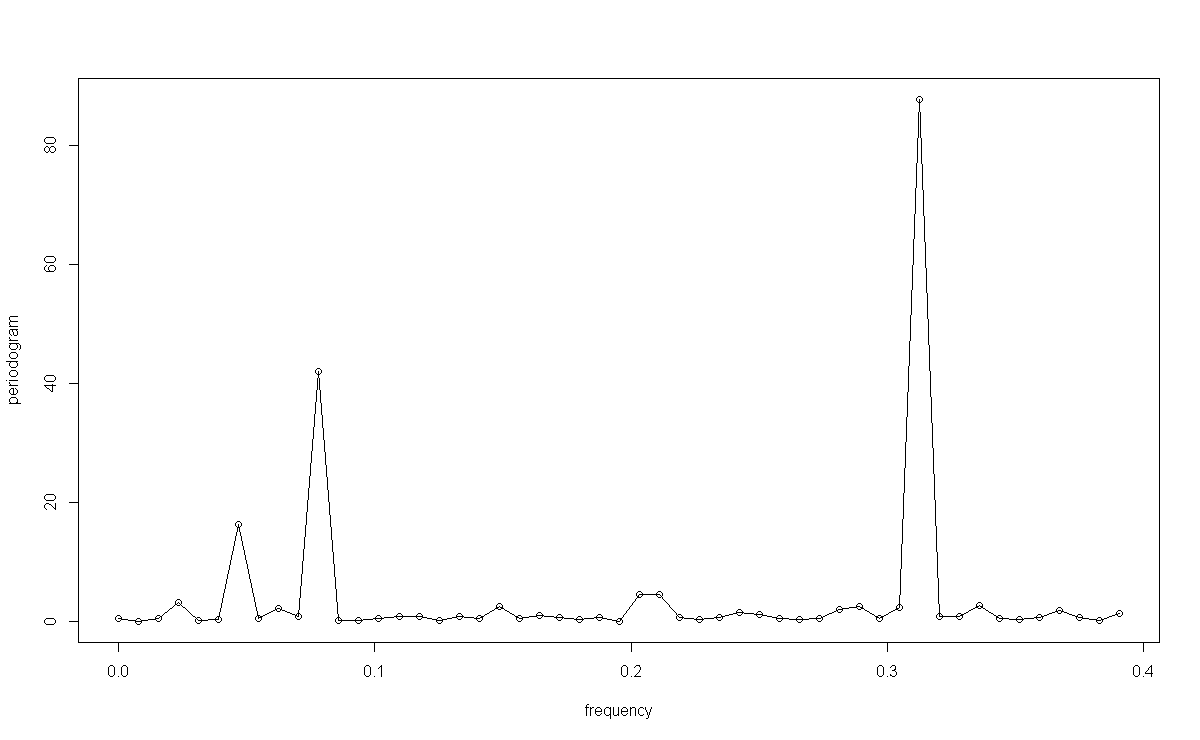
In summary we find that the 128 samples had the effect of moving the predominant frequencies slightly from the original of 0.4, 0.1 and 0.06

Part c. Give the plot of the sum and its periodogram. And, describe the locations of the main peaks in the periodogram.

The sum with and without random is plotted below:



The periodogram of the sum with random noise:



The peaks are at: 0.3125000, 0.0781250, 0.0468750

This can be seen in the following output:

> f

[1] 0.0000000 0.0078125 0.0156250 0.0234375 0.0312500 0.0390625 **0.0468750**

[8] 0.0546875 0.0625000 0.0703125 **0.0781250** 0.0859375 0.0937500 0.1015625

[15] 0.1093750 0.1171875 0.1250000 0.1328125 0.1406250 0.1484375 0.1562500

[22] 0.1640625 0.1718750 0.1796875 0.1875000 0.1953125 0.2031250 0.2109375

[29] 0.2187500 0.2265625 0.2343750 0.2421875 0.2500000 0.2578125 0.2656250

[36] 0.2734375 0.2812500 0.2890625 0.2968750 0.3046875 **0.3125000** 0.3203125

[43] 0.3281250 0.3359375 0.3437500 0.3515625 0.3593750 0.3671875 0.3750000

[50] 0.3828125 0.3906250

> round(P[1:51],7)

[1] 0.5059187 0.0367471 0.4931214 3.1593487 0.0973737 0.3015556 **16.3164595**

[8] 0.4885259 2.1124190 0.7729821 **41.9597151** 0.1278943 0.1962052 0.4019918

[15] 0.7240451 0.8546352 0.1277321 0.8692070 0.4151008 2.4839419 0.4459537

[22] 1.0571380 0.5651252 0.3341213 0.6505136 0.0153921 4.4270196 4.5131360

[29] 0.6014816 0.2596037 0.6470128 1.5140902 1.1810165 0.4544257 0.3847988

[36] 0.4624755 1.9402618 2.4851796 0.4804252 2.2882249 **87.7015372** 0.8442601

[43] 0.8653435 2.7019746 0.4899224 0.3449516 0.6084828 1.8495165 0.6987285

[50] 0.1266436 1.2900805

**3.** Use the dataset “speech.dat” from the Week 6 folder. The data are described in Example 1.3 on page 5 of the textbook. The series is a sample of recorded speech of the phrase aaa…hhh, sampled for 0.1020 seconds at the rate of 10,000 points per second. Thus there are n = 1020 data values. The authors don’t say what exactly is the response variable, but a likely possibility for the response is a measure of the audio frequency of the sound. (Maybe they held out from saying that to avoid confusion with the frequency of a cosine wave.)

The data can be read as x = scan("speech.dat") once the data file is in place on your computer.

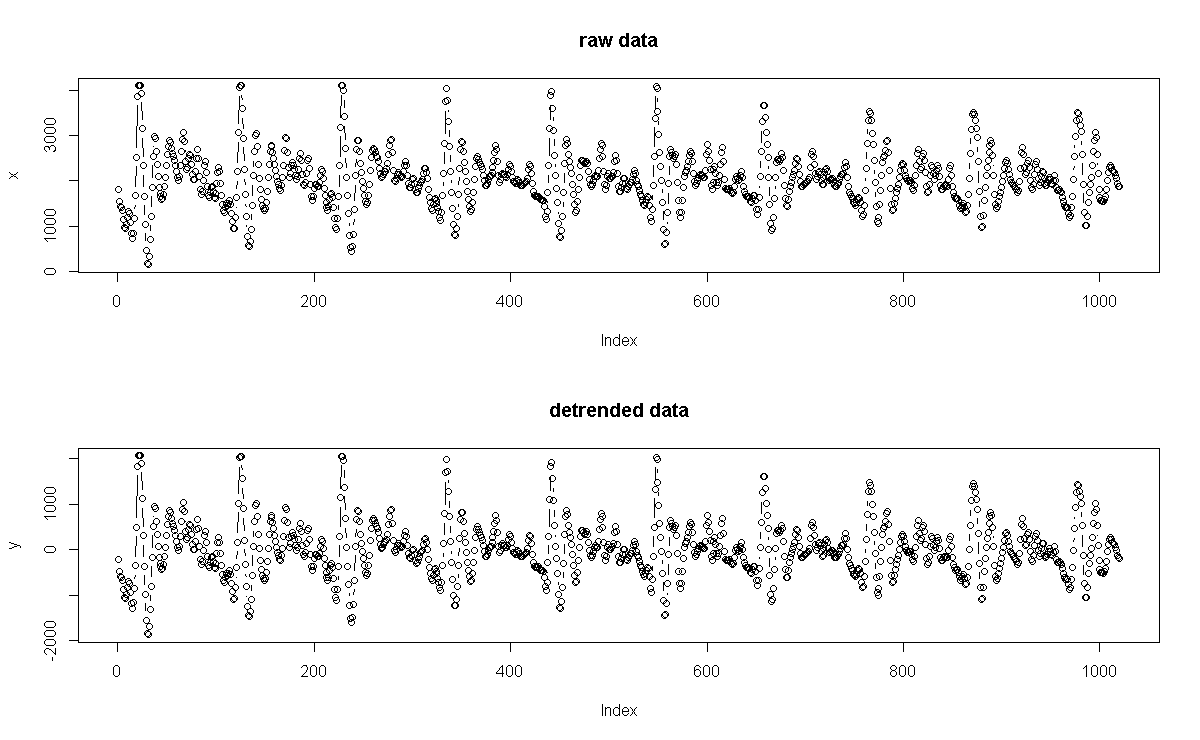
There’s no obvious trend seen in Figure 1.3 on page 6, but it never hurts to detrend before looking at a periodogram. At worst, detrending does nothing. Detrend in R by doing a linear regression on the time index. The detrended values are the residuals from this regression. The sequence of commands might be something like

t = 1:1020

regr = lm(x~t)

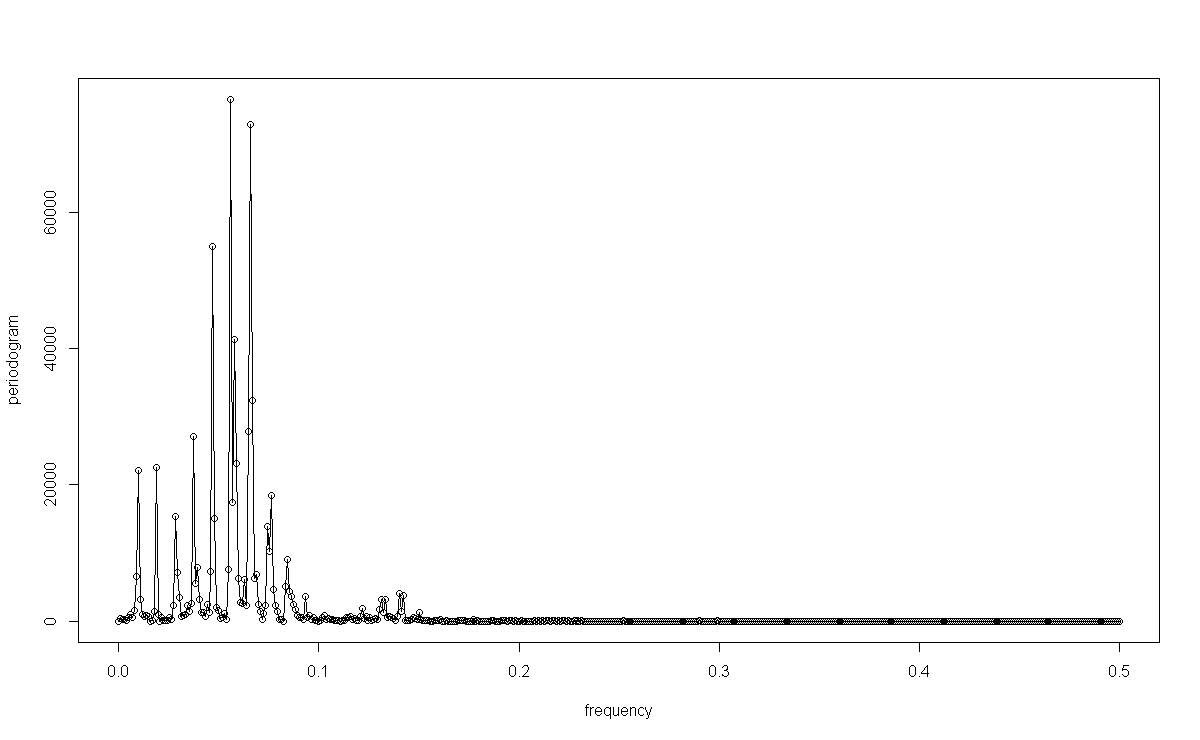
y = residuals(regr)

In the above code, y would be the detrended series, the residuals from a linear regression.



Plot the periodogram of the detrended series. Give the plot and describe the pattern of peaks as they relate to frequency (the horizontal axis).

The periodogram is:

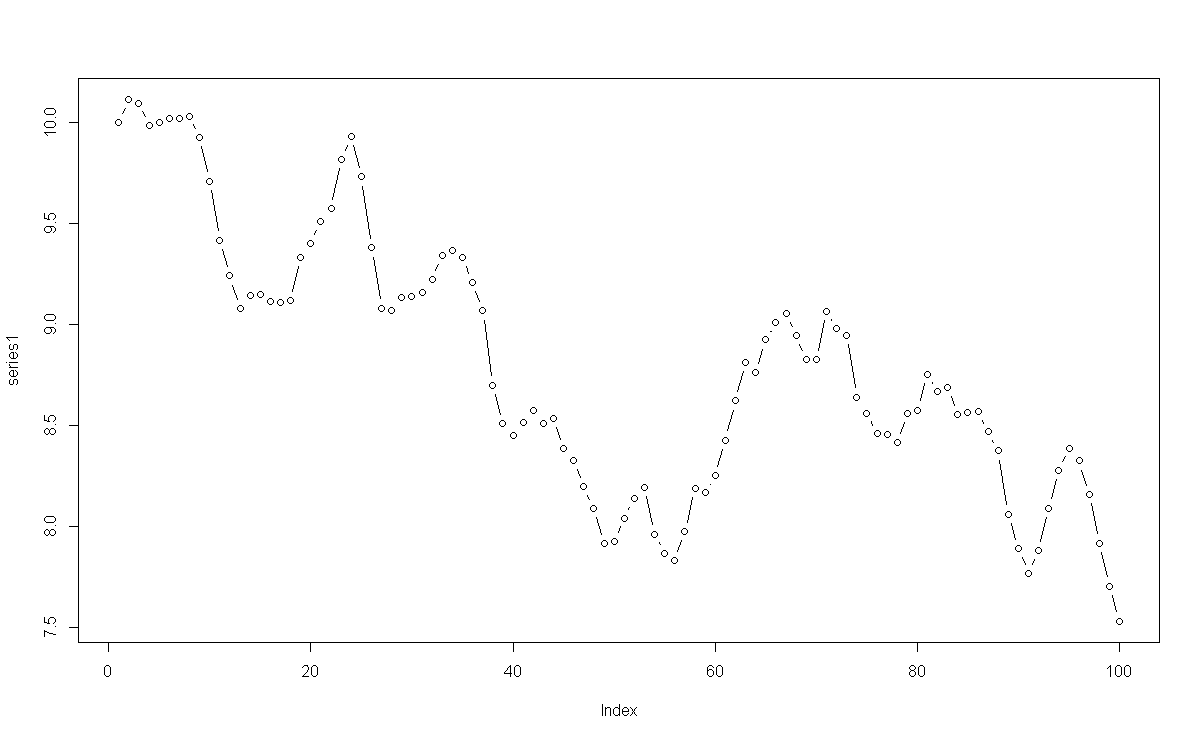


In the above periodogram we see that we have numerous peaking frequencies in the range between 0 and 0.1 followed by a few small peaks aroung 0.14

So we can say that we see many frequencies concentrated in the 0 to 0.1 range and this is something we can notice in the raw data namely the rather regular repetition of small wavelets.

**4.** This problem is a review of the material in Week 3 (non-seasonal ARIMA models). Use the lakesim100.dat dataset in the Week 6 folder. It gives *n* =100 consecutive annual measures of the level of a lake.

a. Plot the time series. Give the plot and an interpretation of the plot as the answer to this part. Please review Lesson 1.1 for what you should interpret in the plot.



Some features of the plot:

- There is a clear downward trend. Since there is a downward linear trend, a first difference may be needed.

- There are no obvious outliers.

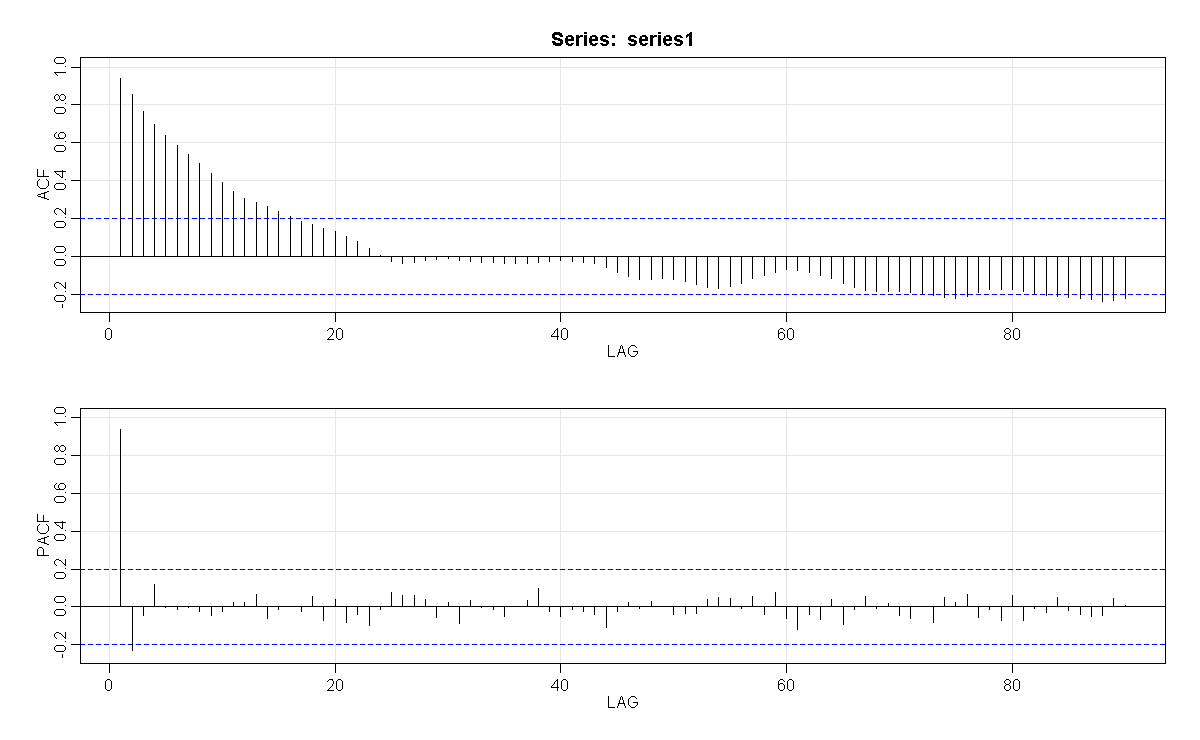
- The variance appears to be constant though it seems difficult to judge with conviction.

- We have annual data and therefore deduce that there is no seasonality

b. Determine the ACF and PACF for the series. (Hint: The most efficient way to do this is to load the astsa library and then use the acf2 command.)

Give the ACF and PACF plots AND write a brief interpretation. Do you think the data are stationary? Do you think you should suggest models from the ACF and PACF?

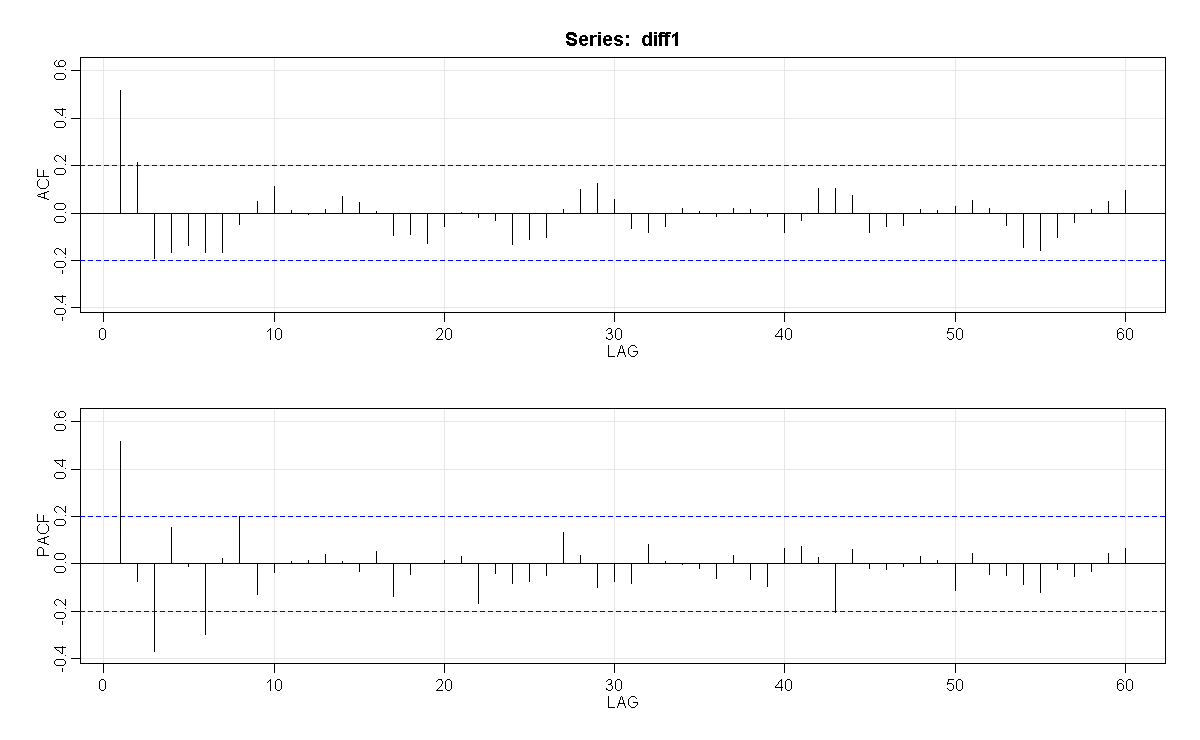
The plots are:



* The PACF points to an AR(2) model but the ACF for an AR(2) model has a sinusoidal behavior which we do observe here.

We did see a downward trend in the raw data and therefore should investigate the difference and the ACF / PACF of the difference before settling on the choice of a model.

c. Give the ACF and PACF of first differences. Write a brief interpretation that indicates what model(s) you think are suggested.



Here we see that:

* ACF has 2 significant peaks and then cuts off – pointing to an MA(2) model. PACF is tailing off.

**While MA(2) seems to be the most convincing model**, we can add a few more possibilities to compare the models.

* Another not very convincing way to think is that the PACF peaks at lag 1 and the ACF peaks at lags 1 and 2, we can try ARMA(1, 2) model. Since we have a diff this translates into ARIMA(1,1,2)
* Since the 2nd peak only slightly touches the significant line, we can also test an MA(1) model.

So we have the following potential models to evaluate:

* ARIMA(0, 1, 2)
* ARIMA(1, 1, 2)
* ARIMA(0, 1, 1)

d. Determine an ARIMA model that you think is suitable for these data. Describe the model and give the estimated coefficients. (You can just give the output that you get from R.)

Let’s compare the models we selected above:

|  |  |
| --- | --- |
| **Model** | **Notes** |
| ARIMA(0, 1, 2) | sigma^2 estimated as 0.009778  $ttable  Estimate SE t.value p.value  ma1 0.7471 0.0690 10.8271 0.0000  ma2 0.8567 0.0648 13.2297 0.0000  constant -0.0218 0.0256 -0.8502 0.3973  $AIC  [1] -3.567581  $AICc  [1] -3.54337  $BIC  [1] -4.489426 |
| ARIMA(1, 1, 2) | sigma^2 estimated as 0.009755  $ttable  Estimate SE t.value p.value  ar1 -0.0571 0.1279 -0.4462 0.6564  ma1 0.7699 0.0844 9.1272 0.0000  ma2 0.8646 0.0665 13.0105 0.0000  constant -0.0220 0.0245 -0.8951 0.3730  $AIC  [1] -3.549944  $AICc  [1] -3.523561  $BIC  [1] -4.445737  We find ar(1) is not significant – leads back to MA(2) |
| ARIMA(0, 1, 1) | sigma^2 estimated as 0.01547  $ttable  Estimate SE t.value p.value  ma1 0.3780 0.0686 5.5069 0.0000  constant -0.0249 0.0172 -1.4482 0.1507  $AIC  [1] -3.128742  $AICc  [1] -3.106242  $BIC  [1] -4.076638 |

ARIMA(0, 1, 2) gives us the best parameters and we choose that as our model of preference.

For ARIMA(0, 1, 2) the below output gives the coefficients:

$ttable

Estimate SE t.value p.value

ma1 0.7471 0.0690 10.8271 0.0000

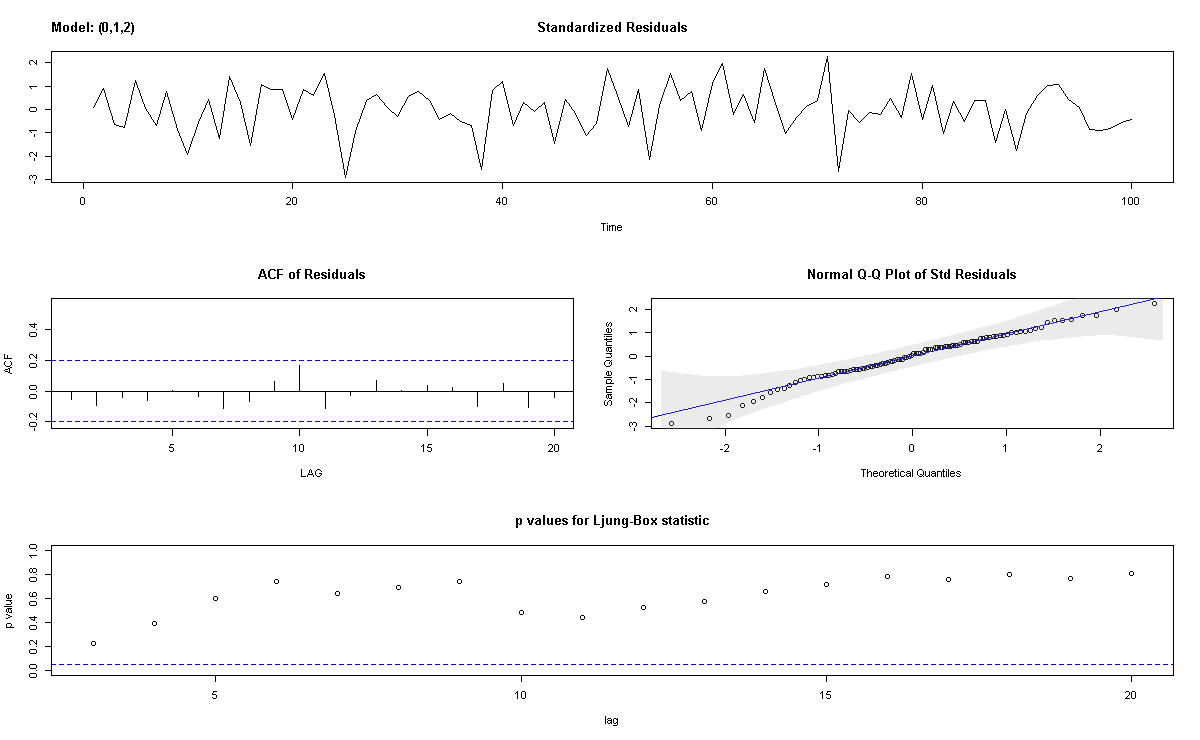
ma2 0.8567 0.0648 13.2297 0.0000

constant -0.0218 0.0256 -0.8502 0.3973

We can see that both ma1 and ma2 are significant.

e. For your model, give diagnostic plots/tests for checking the suitability of the model (for example, the diagnostic output from the sarima command). Briefly discuss what the plots indicate about your model.

The diagnostic plots are:



Based on this we deduce the following:

* The time series plot of the standardized residuals mostly indicates that there’s no trend in the residuals, no outliers, and in general, no changing variance across time.
* The ACF of the residuals doesn’t show any significant autocorrelations– a good result overall.
* The Q-Q plot is a normal probability plot – The plot looks almost linear except in the lower left side - the assumption of normally distributed residuals more or less holds though not completely convincingly.
* The bottom plot gives p-values for the Ljung-Box-Pierce statistics for each lag up to 20. All p-values are above the dashed blue line - That’s a good result.